Project Proposal: Advanced Biomedical Extractive Question Answering System

1. Introduction

Biomedical Extractive Question Answering (BEQA) focuses on retrieving precise answers from biomedical literature, playing a crucial role in information retrieval and decision-making within healthcare and life sciences. Due to the vast volume of biomedical texts, there is an increasing demand for systems that can process complex data accurately and efficiently. However, existing BEQA systems struggle with the intricacies of domain-specific terminology and require fine-tuning for specific question types, such as factoid queries.

2. Background and Motivation

Recent advancements in transformer models (e.g., BERT, BioBERT, ClinicalBERT) have vastly improved natural language processing (NLP) applications. However, BEQA systems still face significant challenges, especially with questions that require precise answers. This project aims to bridge the gap by developing a system optimized for handling factoid-type biomedical questions.

3. Project Objectives

The primary objectives of the project are:

- **Develop and Fine-tune Models**: Fine-tune six transformer models (BERT, BioBERT, PubMedBERT, ClinicalBERT, Bio+ClinicalBERT, and BioLinkBERT) on the BioASQ factoid dataset to extract precise answers.
- **Enhance Entity Recognition and Relationship Modeling**: Integrate Named Entity Recognition (NER) and Relation Extraction (RE) techniques to identify biomedical entities and their relationships.
- **Integrate BiLSTM Layers**: Add Bidirectional Long Short-Term Memory (BiLSTM) layers to capture contextual and sequential information more effectively.
- **Optimize Model Performance**: Implement ensemble learning techniques (e.g., bootstrap sampling, majority voting) to enhance the accuracy and reliability of the BEQA system.

4. Methodology

Model Development and Fine-tuning

The project involves fine-tuning six pre-trained models (BERT, BioBERT, PubMedBERT, ClinicalBERT, Bio+ClinicalBERT, and BioLinkBERT) using the BioASQ factoid dataset. The models will be evaluated using F1 and Exact Match (EM) scores to measure performance, with Bio+ClinicalBERT expected to perform best due to its comprehensive training on both biomedical and clinical texts.

Incorporation of NER and RE Techniques

NER will identify critical biomedical entities such as diseases, proteins, and drugs, while RE will model relationships between these entities, improving the model's context comprehension.

Addition of BiLSTM Layers

Adding BiLSTM layers will help the model capture long-range dependencies and improve understanding of the biomedical texts, especially in lengthy passages.

Ensemble Learning

The output of multiple models will be combined using bootstrap sampling and majority voting. This will create a more robust system capable of producing reliable answers.

5. Evaluation Metrics and Expected Results

- **F1 Score**: Expected to improve with fine-tuning and ensemble techniques.
- **Exact Match (EM)**: Anticipated to be higher than baseline models.
- **Comprehension**: Improved ability to capture complex biomedical relationships.
- **Reliability**: Enhanced system reliability due to ensemble learning methods.

6. Expected Impact

This project aims to create a state-of-the-art BEQA system that improves upon existing models by offering better precision and reliability. It will also provide a framework for future research in applying advanced NLP techniques, such as transformer models and BiLSTM, to domain-specific question-answering tasks. The

system's ability to support accurate information retrieval could have a significant impact on healthcare decision-making processes, potentially aiding clinicians and researchers in accessing vital information.

7. Conclusion

This project sets out to push the boundaries of Biomedical Extractive Question Answering (BEQA) systems by utilizing cutting-edge transformer-based models, enhanced with NER, RE, and BiLSTM layers. The system has the potential to set new benchmarks in BEQA performance, providing a reliable and efficient method for extracting answers from complex biomedical texts.